Les discussions Wikipedia : un corpus pour caractériser le genre "(wiki) discussion"

Lydia-Mai Ho-Dac et Véronika Laippala

CLLE-ERSS, TIAS

CMO, 24-25 octobre, Rennes

Plan

Plan

Problème du Web as Corpus : une (trop) vaste variété de "genres"

Enjeux : définir les *genres* du web

- Profiler les textes
- Comprendre la distribution des genres différents

⇒ Développer des méthodes quantitatives pour l'analyse et l'identification des genres du web

La notion de genre

Un genre implique une communauté et une visée discursive plus ou moins délimitées

> *By means the concept of genre we can approach texts from the macro-level as communicative acts within a discourse network or system (Trosborg 1997 :7)*

> *le genre est une « catégorie de textes fondée sur une pratique sociale établie, définie a priori. La catégorie est reconnue et validée par le fait qu'elle*

Problème du Web as Corpus : une (trop) vaste variété de "genres"

Enjeux : définir les *genres* du web

- Profiler les textes
- Comprendre la distribution des genres différents

$\Rightarrow$ Développer des méthodes quantitatives pour l'analyse et l'identification des genres du web

Proposition méthodologique

1. Sélectionner une sous-parties du Web au genre mieux défini : le monde la Wikipedia
2. Décrire les traits typiques des genres représentés
3. Classer les textes du web selon ces genres, en se basant sur ces traits typiques

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Des genres et situations de communication plutôt bien définies

- articles encyclopédiques
- discussions

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Des genres et situations de communication plutôt bien définies

- articles encyclopédiques
- discussions
  - autour de la rédaction collaborative d'un article
  - autour du projet Wikipedia («cafés et bistrots», e.g. «Forum des Nouveaux», «Le salon de médiation», etc.)

Wikipedia As Corpus

Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Des genres et situations de communication plutôt bien définies

- articles encyclopédiques
- discussions
  - autour de la rédaction collaborative d'un article
  - autour du projet Wikipedia («cafés et bistrots», e.g. «Forum des Nouveaux», «Le salon de médiation», etc.)
- Journaux d'activité («Bulletin des patrouilleurs»)
- Ateliers et "Machines à café"

Wikipedia As Corpus

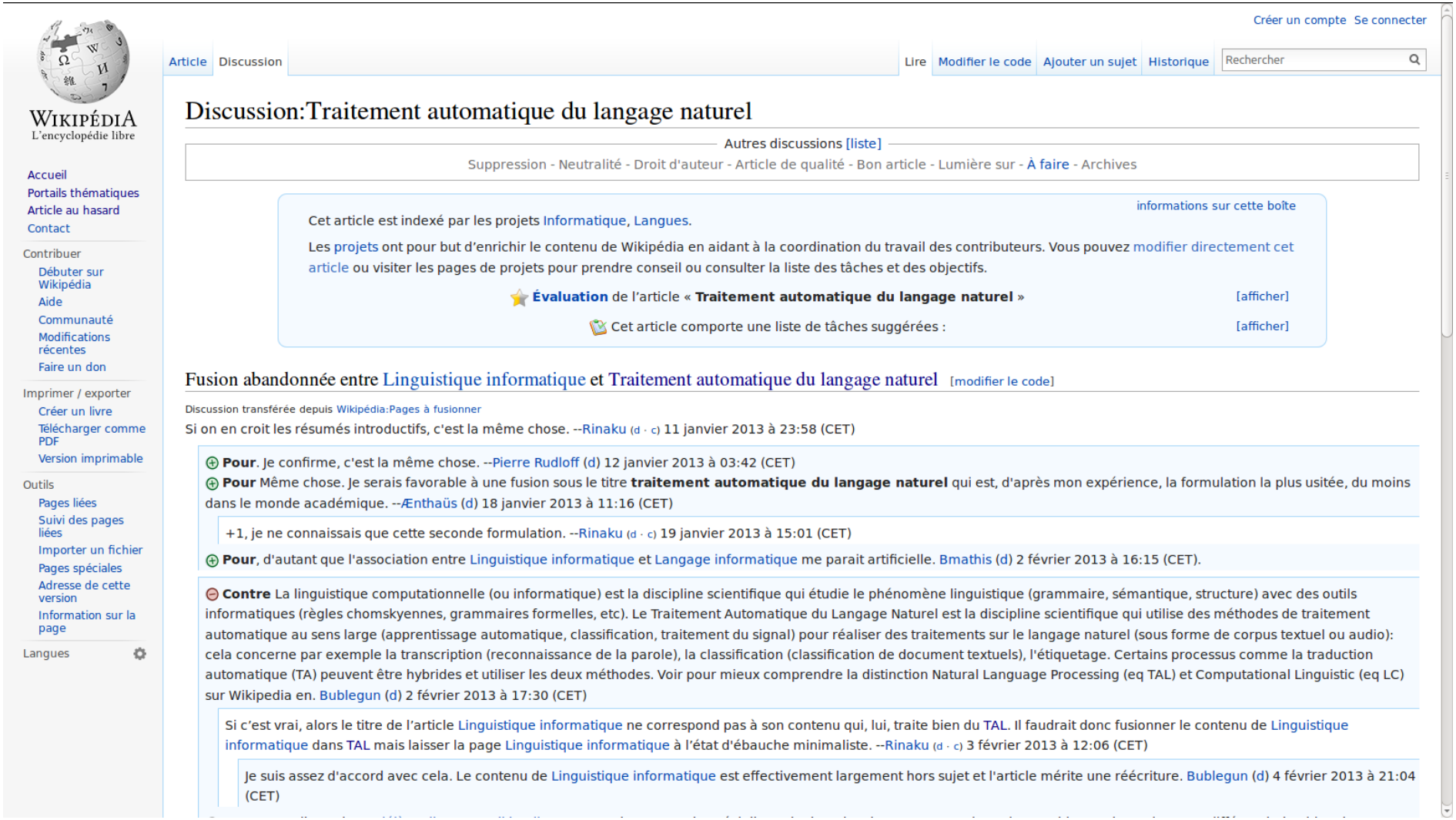
Accessibilité et quantité des données

- contenu libre distribué publiquement (Creative Commons by-sa)
- depuis 2001
- existe dans presque toutes les langues → objet d'étude international

Des genres et situations de communication plutôt bien définies

- articles encyclopédiques
- **discussions**
  - autour de la rédaction collaborative d'un article
  - autour du projet Wikipedia («cafés et bistrots», e.g. «Forum des Nouveaux», «Le salon de médiation», etc.)
- Journaux d'activité («Bulletin des patrouilleurs»)
- Ateliers et "Machines à café"

Les discussions Wikipedia



WIKIPÉDIA
L'encyclopédie libre

Créer un compte Se connecter

Article | Discussion | Lire | Modifier le code | Ajouter un sujet | Historique | Rechercher

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Imprimer / exporter
Créer un livre
Télécharger comme PDF
Version imprimable

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Adresse de cette version
Information sur la page

Langues

Discussion:Traitement automatique du langage naturel

Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

informations sur cette boîte

Cet article est indexé par les projets Informatique, Langues.

Les projets ont pour but d'enrichir le contenu de Wikipédia en aidant à la coordination du travail des contributeurs. Vous pouvez modifier directement cet article ou visiter les pages de projets pour prendre conseil ou consulter la liste des tâches et des objectifs.

Évaluation de l'article « **Traitement automatique du langage naturel** » [afficher]

Cet article comporte une liste de tâches suggérées : [afficher]

Fusion abandonnée entre Linguistique informatique et Traitement automatique du langage naturel [modifier le code]

Discussion transférée depuis Wikipédia:Pages à fusionner

Si on en croit les résumés introductifs, c'est la même chose. --Rinaku (d · c) 11 janvier 2013 à 23:58 (CET)

⊕ **Pour**. Je confirme, c'est la même chose. --Pierre Rudloff (d) 12 janvier 2013 à 03:42 (CET)

⊕ **Pour** Même chose. Je serais favorable à une fusion sous le titre **traitement automatique du langage naturel** qui est, d'après mon expérience, la formulation la plus usitée, du moins dans le monde académique. --Ænthaüs (d) 18 janvier 2013 à 11:16 (CET)

+1, je ne connaissais que cette seconde formulation. --Rinaku (d · c) 19 janvier 2013 à 15:01 (CET)

⊕ **Pour**, d'autant que l'association entre Linguistique informatique et Langage informatique me parait artificielle. Bmathis (d) 2 février 2013 à 16:15 (CET).

⊖ **Contre** La linguistique computationnelle (ou informatique) est la discipline scientifique qui étudie le phénomène linguistique (grammaire, sémantique, structure) avec des outils informatiques (règles chomskyennes, grammaires formelles, etc). Le Traitement Automatique du Langage Naturel est la discipline scientifique qui utilise des méthodes de traitement automatique au sens large (apprentissage automatique, classification, traitement du signal) pour réaliser des traitements sur le langage naturel (sous forme de corpus textuel ou audio): cela concerne par exemple la transcription (reconnaissance de la parole), la classification (classification de document textuels), l'étiquetage. Certains processus comme la traduction automatique (TA) peuvent être hybrides et utiliser les deux méthodes. Voir pour mieux comprendre la distinction Natural Language Processing (eq TAL) et Computational Linguistic (eq LC) sur Wikipedia en. Bublegun (d) 2 février 2013 à 17:30 (CET)

Si c'est vrai, alors le titre de l'article Linguistique informatique ne correspond pas à son contenu qui, lui, traite bien du TAL. Il faudrait donc fusionner le contenu de Linguistique informatique dans TAL mais laisser la page Linguistique informatique à l'état d'ébauche minimaliste. --Rinaku (d · c) 3 février 2013 à 12:06 (CET)

Je suis assez d'accord avec cela. Le contenu de Linguistique informatique est effectivement largement hors sujet et l'article mérite une réécriture. Bublegun (d) 4 février 2013 à 21:04 (CET)

Les discussions Wikipedia

Accessibilité et quantité des données

- (sorte de) Forum de discussion libre distribué publiquement (Creative Commons by-sa)
- existe dans une multitude de langues $\rightarrow$ objet d'étude international

Richesse des métadonnées

- thématique (portail thématique, article associé)
- connaissances partagées (article associé)
- indice de subjectivité (appel au calme, etc.)
- informations sur le locuteur (statut dans la communauté, participation à la Wikipedia, possibilité de profilage)

Plan

Procédure de constitution

1. Extraction des discussions depuis le dump

   sauvegarde globale des pages courantes de la Wikipedia française (archive frwiki-20150512-pages-meta-current#.xml.bz2 diffusée librement sur la page http ://dumps.wikimedia.org/frwiki/20150512/)
2. Sélection des discussions "à garder"
3. Analyse des objets textuels constitutifs de chaque discussion : sections (fils), messages
4. Conversion selon la TEI-P5
5. Analyse syntaxique automatique (Talismane, Urieli 2013)

Procédure de constitution - sélection des discussions

/<title>Discussion/ sur le dump du 20150512 : 3 487 480

Discussions portant sur un utilisateur	1 990 927	57%
Discussions portant sur un article	**1 496 553**	**43%**
Discussions redirigées vers une autre discussion	116 432	8%
Discussions vides ou contenant moins de 2 mots	1 013 791	68%
Discussions retenues	**366 326**	**24%**

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. **Extraction des méta-données**
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Extraction des méta-données

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. **Structuration en sections (fils) et messages (posts)**
3. **Délimitation des différentes contributions** :
   - 1 message $\leftrightarrow$ 1 (Signature+) date de publication
   - 1 message 1 $\leftrightarrow$ TEI-P5 <sp> *(speech) An individual speech in a performance text, or a passage presented as such in a prose or verse text.*
4. Évaluation de l'extraction

Procédure de constitution - structuration des discussions

Structuration des discussions

Procédure de constitution - structuration des discussions

Des discussions à la norme TEI-P5

1. Extraction des méta-données
2. Structuration en sections (fils) et messages (posts)
3. Délimitation des différentes contributions
4. **Évaluation de l'extraction**

Évaluation de la constitution

Évaluation de l'extraction

- 7 discussions évaluées manuellement : 413 messages et 47 284 mots
- précision = 0,92, rappel = 0,95

Bruit 3 messages vides ; 5 messages scindés en 2 ; 25 messages fusionnant 2 ou 3 messages

Silence 23 messages absents

Évaluation de la constitution - exemples d'erreur

sens du mot Lehi

66	0	Le truc c'est qu'en hébreu, et en arabe, la frontière est parfois mince entre un acronyme et une abréviation. "Fatah" ou "Hamas" par exemple sont-ils acronymes ou abréviations ? --Markov 5 septembre 2006 à 02:04 (CEST)
67	1	Je pense que ce sont res rétro-acronymes : on prend un mot qui veut dire quelque chose, et on invente un sigle dont l'acronyme deviendra le mot choisit. C'est ça ? Par contre je ne crois pas que Lehi veuille dire quelque chose en Hébreu. Tu comprend l'hébreu, Markov ? Christophe Cagé - liste de mes articles 5 septembre 2006 à 07:06 (CEST)
68	3	Euh, "ksat, ksat", (très peu), notions de base. --Markov 8 septembre 2006 à 11:02 (CEST)
69	2	A ma connaissance Lehi, ne veut rien dire mais je ne suis pas la référence. En cherchant sur google j'ai trouvé que c'était un lieu dit où les Philistins et les hébreux s'affrontèrent mais cela ne prouve pas la volonté de faire le lien... Ceedjee 5 septembre 2006 à 07:45 (CEST)
70	0	Ben si les hébreux ont gagnés, c'est en tout cas un indice. C'est le cas ? Christophe cagé Au delà du renommage, il serait utile de pourvoir faire la distinction entre une abréviation par acronymie ou par un sigle. Personnellement, je n'ai jamais vu d'abréviations écrits en lettres capitales, c'est de cette façon que j'avais pensé que LEHI était un sigle. Maintenant le cas de l'hébreux semblent particulier, je ne suis pas apte à trancher (surtout pour une Rétro-acronymie). VIGNERON * 5 septembre 2006 à 08:56 (CEST)
71	1	Tu a raison. J'ai mis LEHI en majuscule, parceque c'est la graphie de Schattner, mais d'autres historiens mettent des minuscules, je crois. Il faut que je vérifie. Sinon, j'ai demandé son avis à Franckiz. Il a un niveau de base en hébreu (mais ce n'est pas un expert, sauf erreur) Christophe Cagé Non, LEHI ne veut rien dire en hébreu. En tout cas rien qui puisse avoir un rapport avec le contexte. ("Léhi" : "Va" à l'impératif féminin). zeeev 5 septembre 2006 à 15:29 (CEST)
72	1	De plus le rajout du "e" dans l'abréviation vient du fait que l'importance des voyelles en hébreu est secondaire. Ce sont les consomnes seulement qui composent la racine du mot hébreu. On appelle donc cette organisation "Léhi", mais il est vrai qu'en toute logique, on aurait pu l'appeler "Lahi" ou "Lohi".....zeeev 5 septembre 2006 à 15:37 (CEST)
73	2	Ta réponse n'est que partielle. Tu dit "pourquoi pas un e". Certe, mais pourquoi pas LHI ? Ca ne se fait pas, en hébreu ?Christophe Cagé - liste de mes articles 6 septembre 2006 à 06:14 (CEST)

sens du mot Lehi [modifier le code]

Le truc c'est qu'en hébreu, et en arabe, la frontière est parfois mince entre un acronyme et une abréviation. "Fatah" ou "Hamas" par exemple sont-ils acronymes ou abréviations ? --Markov (discut.) 5 septembre 2006 à 02:04 (CEST)

Je pense que ce sont res rétro-acronymes : on prend un mot qui veut dire quelque chose, et on invente un sigle dont l'acronyme deviendra le mot choisit. C'est ça ? Par contre je ne crois pas que Lehi veuille dire quelque chose en Hébreu. Tu comprend l'hébreu, Markov ? Christophe Cagé - liste de mes articles 5 septembre 2006 à 07:06 (CEST)

Euh, *ksat, ksat*, (très peu), notions de base. --Markov (discut.) 8 septembre 2006 à 11:02 (CEST)

A ma connaissance Lehi, ne veut rien dire mais je ne suis pas la référence. En cherchant sur google j'ai trouvé que c'était un lieu dit où les Philistins et les hébreux s'affrontèrent mais cela ne prouve pas la volonté de faire le lien... Ceedjee *contact* 5 septembre 2006 à 07:45 (CEST)

Ben si les hébreux ont gagnés, c'est en tout cas un indice. C'est le cas ? Christophe cagé

Au delà du renommage, il serait utile de pourvoir faire la distinction entre une abréviation par acronymie ou par un sigle. Personnellement, je n'ai jamais vu d'abréviations écrits en lettres capitales, c'est de cette façon que j'avais pensé que LEHI était un sigle. Maintenant le cas de l'hébreux semblent particulier, je ne suis pas apte à trancher (surtout pour une Rétro-acronymie). VIGNERON * discut. 5 septembre 2006 à 08:56 (CEST)

Tu a raison. J'ai mis LEHI en majuscule, parceque c'est la graphie de Schattner, mais d'autres historiens mettent des minuscules, je crois. Il faut que je vérifie. Sinon, j'ai demandé son avis à Franckiz. Il a un niveau de base en hébreu (mais ce n'est pas un expert, sauf erreur) Christophe Cagé

Non, LEHI ne veut rien dire en hébreu. En tout cas rien qui puisse avoir un rapport avec le contexte. ("Léhi" : "Va" à l'impératif féminin). zeeev 5 septembre 2006 à 15:29 (CEST)

De plus le rajout du "e" dans l'abréviation vient du fait que l'importance des voyelles en hébreu est secondaire. Ce sont les consomnes seulement qui composent la racine du mot hébreu. On appelle donc cette organisation "Léhi", mais il est vrai qu'en toute logique, on aurait pu l'appeler "Lahi" ou "Lohi".....zeeev 5 septembre 2006 à 15:37 (CEST)

Ta réponse n'est que partielle. Tu dit "pourquoi pas un e". Certe, mais pourquoi pas LHI ? Ca ne se

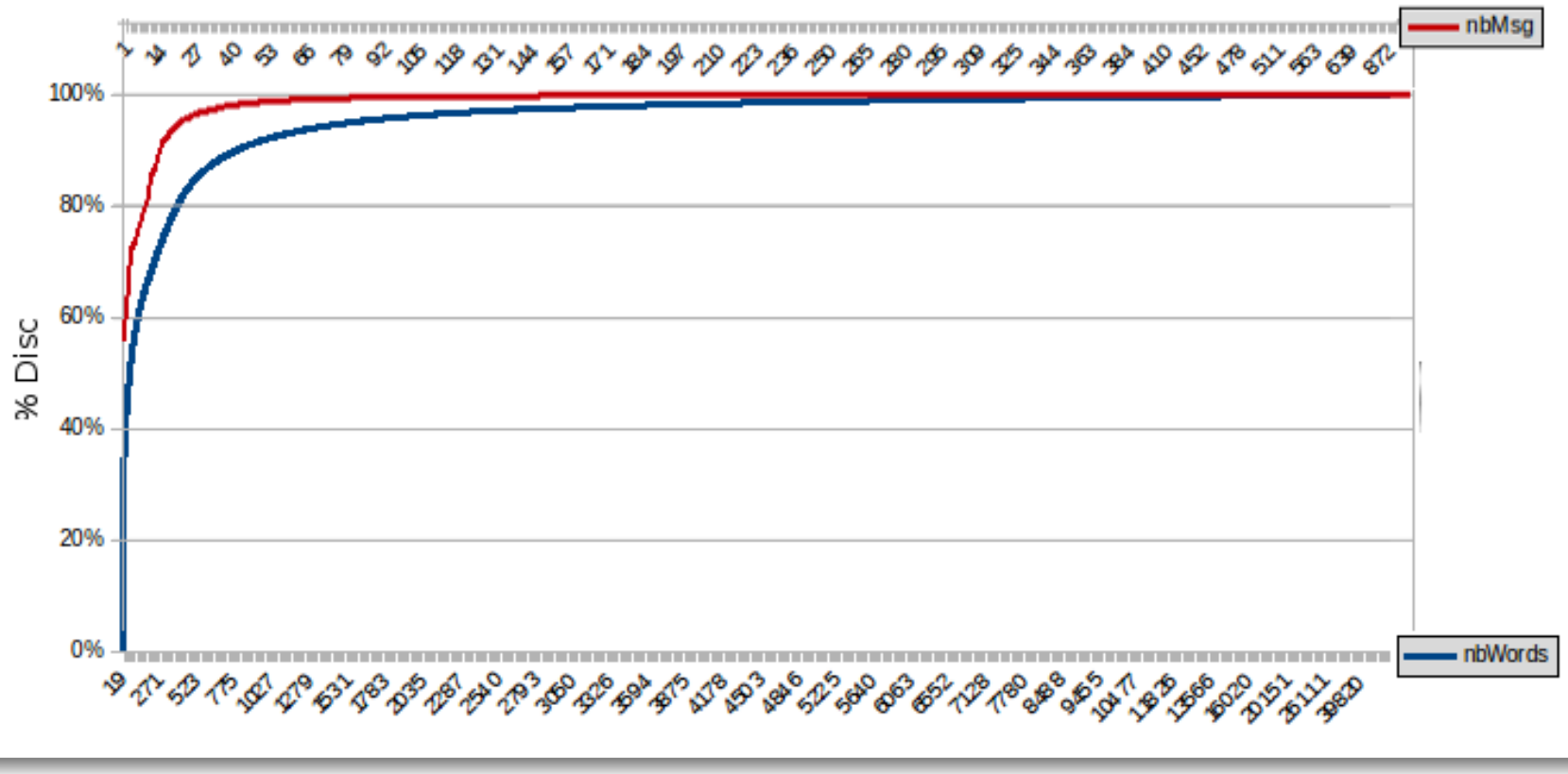
Corpus constitué et bientôt mis à disposition

discussions	sections	messages	mots
366 326	1 024 351	3 022 240	159 578 279

Disponible très prochainement sur `http://redac.univ-tlse2.fr/`

Caractéristiques globales - longueur

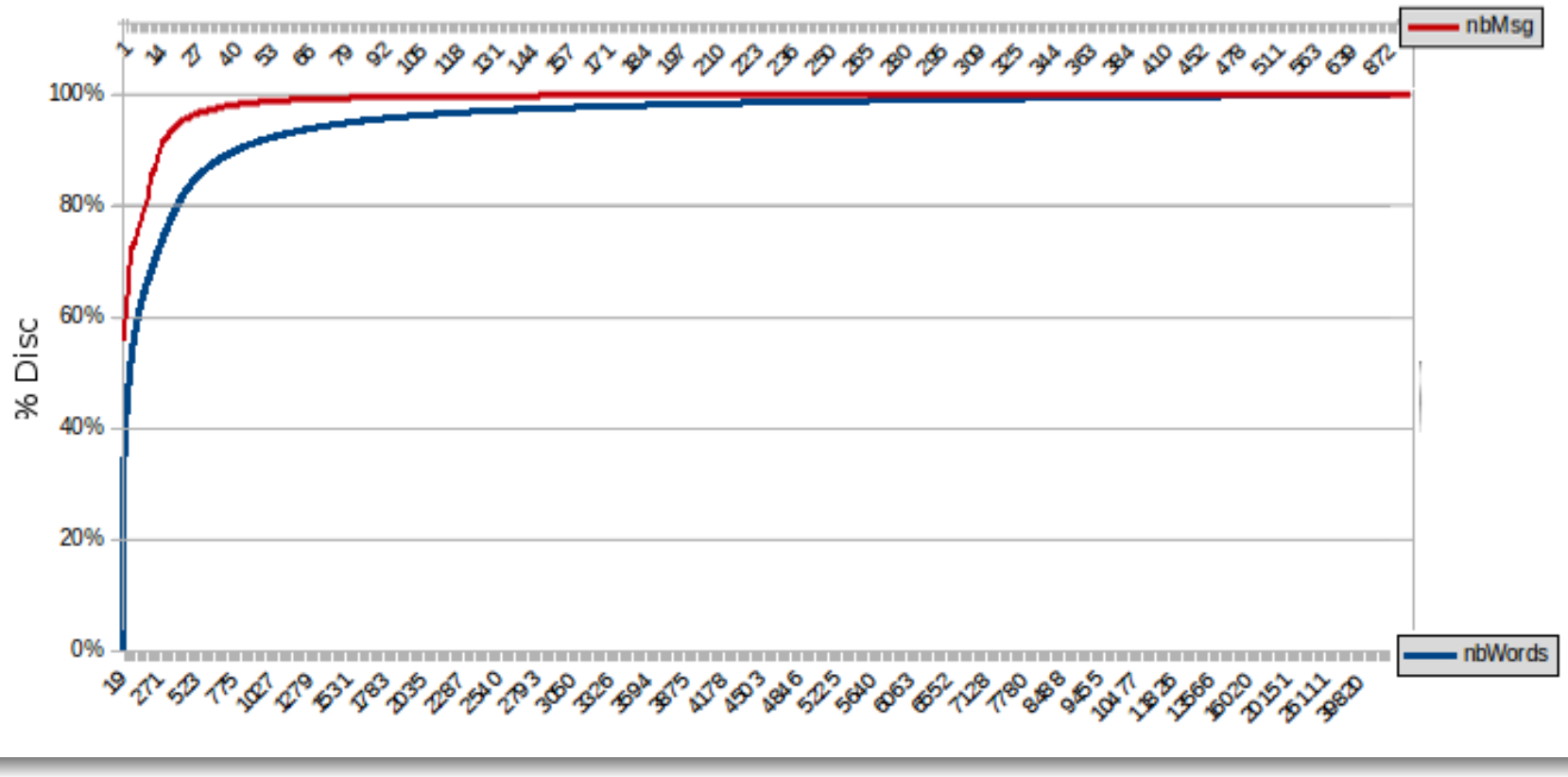
202 856 (55%) discussions ne contenant qu'un message
181 503 (50%) contenant moins de 53 mots



Caractéristiques globales - longueur

Des discussions allant jusqu'à 1 143 messages et 148 968 mots

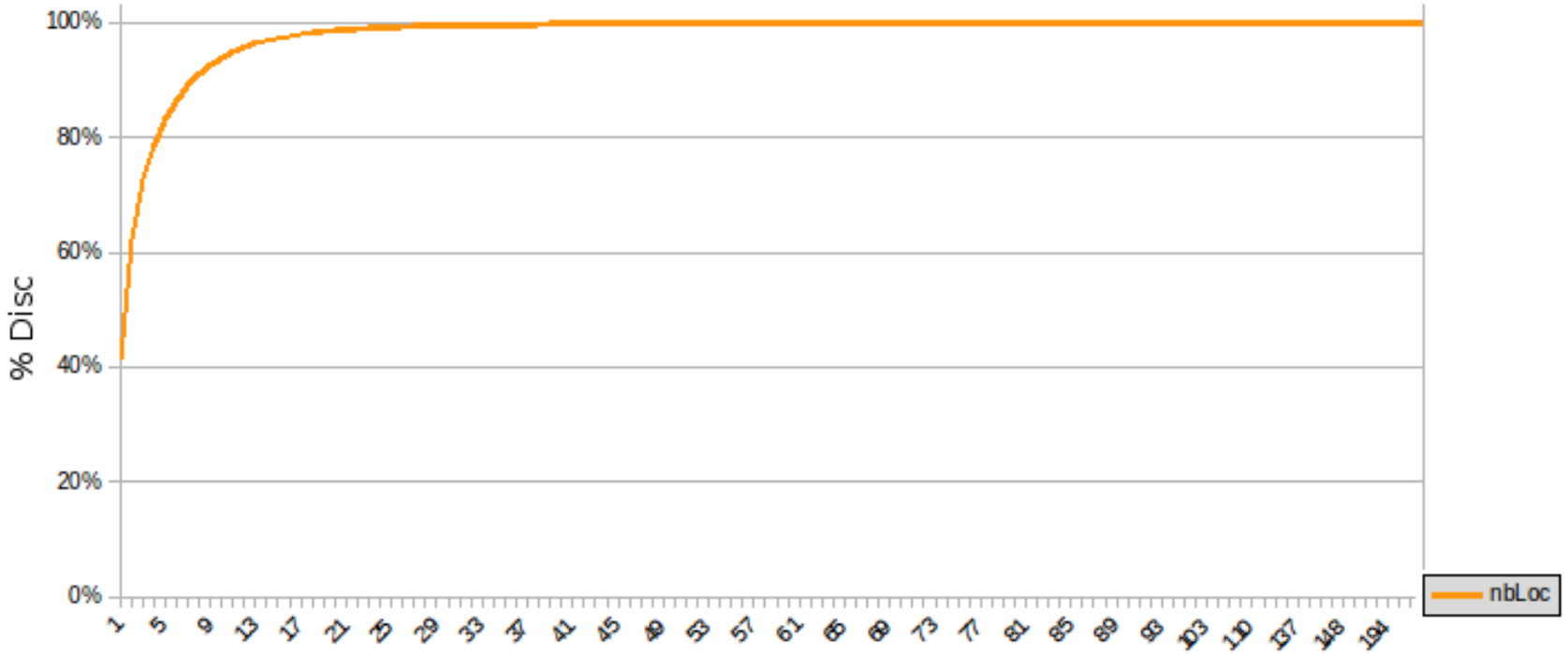
"*Opposition au mariage homosexuel en France part1*"



Caractéristiques globales - interactions

150 603 (41%) monologues
40 413 (10%) impliquant entre 8 et 228 locuteurs différents*

"*Opposition au mariage homosexuel en France part1*" : 93 locuteurs (anonyme = 1)



80% de contributions anonymes

* Discussion sur l'admissibilité de la page "Mickaël Vendetta"

Plan

Étude contrastive de plusieurs genres pour **caractériser** le genre "discussion"

Approche *hypothesis-driven* : mesurer des caractéristiques *a priori* spécifiques

Degré de "déviance" langagière, traces de subjectivité, structures discursives / interactions (formules d'ouverture)

Approche data-driven : découvrir les caractéristiques des corpus

Poids de N-grams lexicaux, morphologiques, syntaxiques dans une tâche classification automatique supervisée

Corpus de comparaison

Constitués

Corpus écrits	No tokens	Description
Rue89	2 192 995	Presse en ligne
AgoraVox	4 099 662	Media citoyen
Forum Santé	236 368 151*	Forum de discussion
		2 585 188 messages
WikiDiscussion (2015)	132 406 816*	Forum de discussion
		3 022 240 messages
WikiArticles (2013)	226 207 672*	Articles encyclopédiques

Parsés

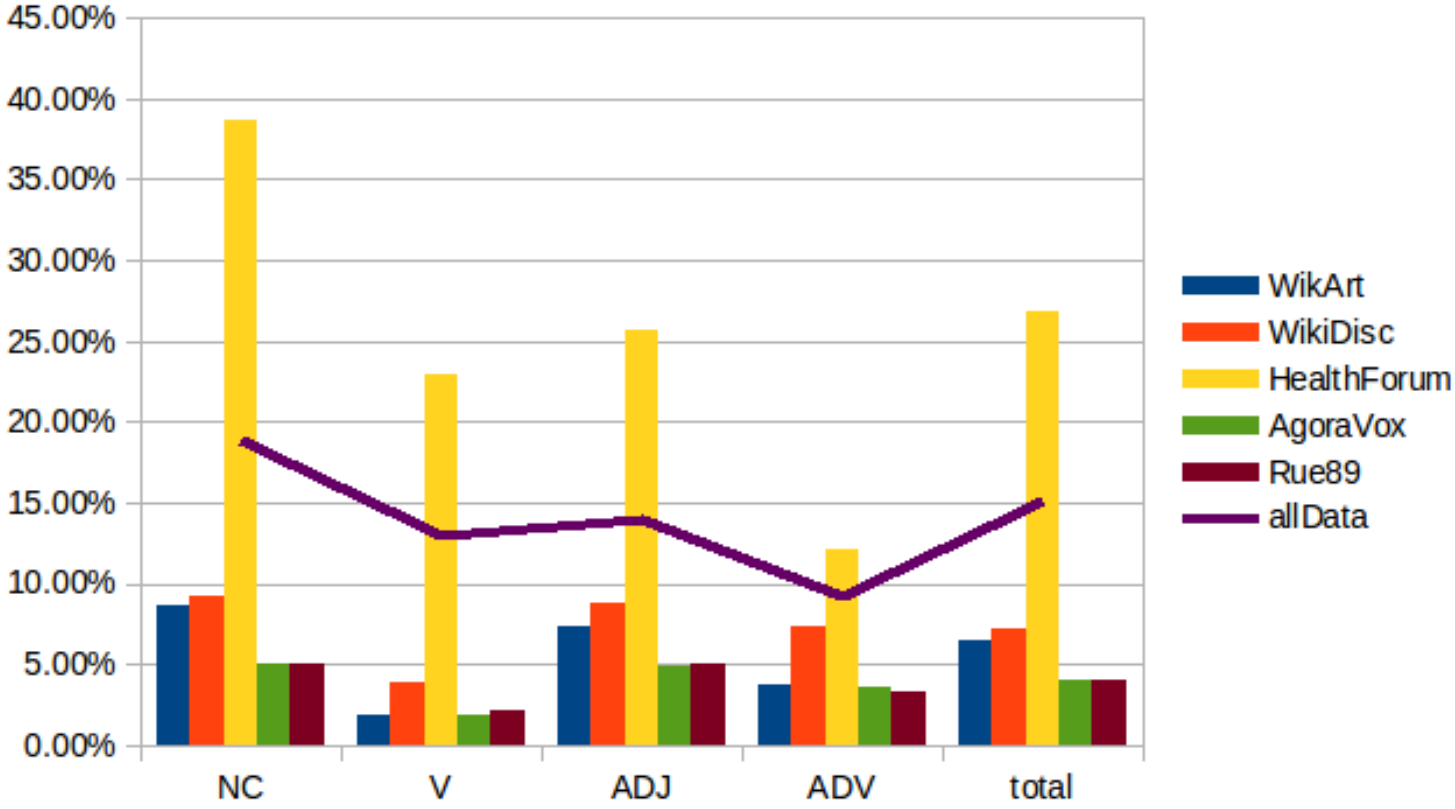
Tous les corpus ont été analysés syntaxiquement avec Talismane (version 1.8.5b, beam de 5, modèle svn)

* Analyses data-driven effectuées sur un sous-ensemble (temps de traitement +++ !)

Évaluation du degré de "déviance" de l'écriture

Méthode : mesurer le taux de mots inconnus

% de mots inconnus selon Talismane (FTB et Lefff) dans les catégories nom, verbe, adjectif et adverbe



Traces de subjectivité

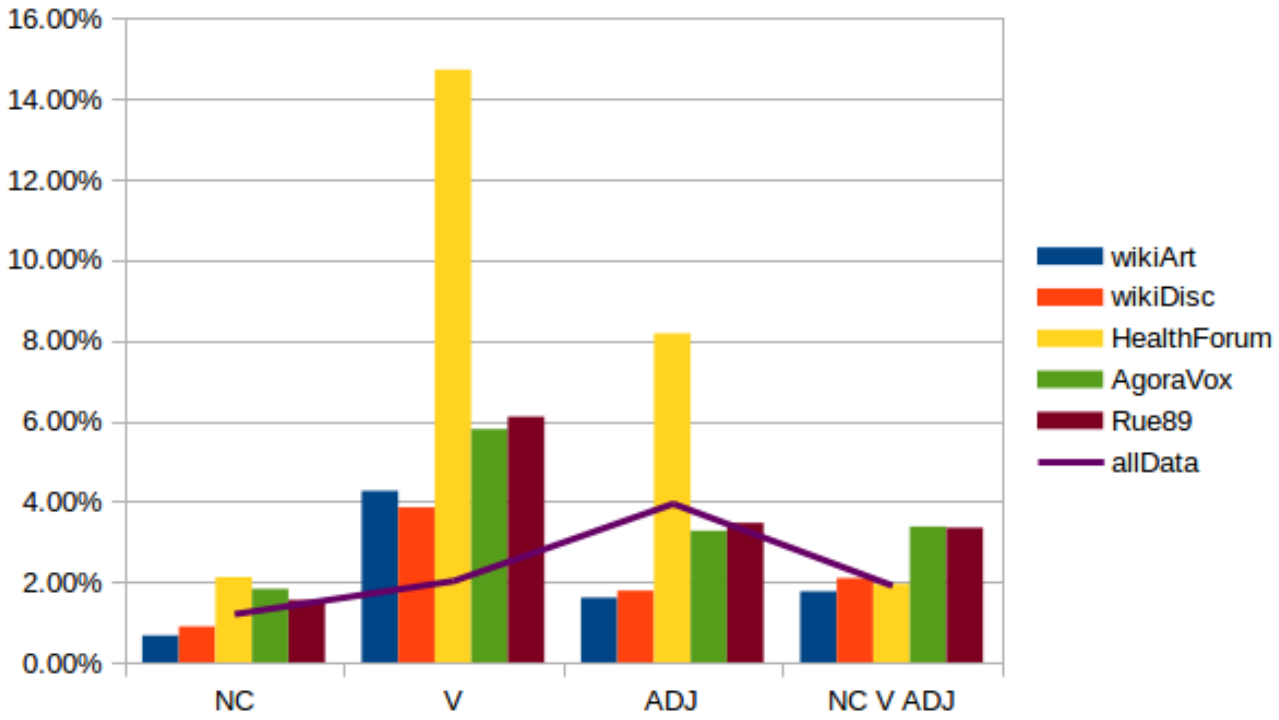
Méthode : projection d'un lexique des affects (Augustyn et al. 2008)

Noms	Verbes	Adjectifs
272	394	493
gratitude	épater	chouette
joie	étonner	beau
exaspération	énerver	agaçant
amertume	emmerder	chiant
choc	interdire	infect
peur	demander	confus
plaisir	gêner	épaté
...	...	...

Traces de subjectivité

Méthode : projection d'un lexique des affects (Augustyn et al. 2008)

% de Noms, Verbes et Adjectifs mentionnés dans le lexique des affect



croire	*67964*
demander	
attendre	
considérer	
doute	
aimer	
respecter	
garder	
définir	
souhaiter	
démontrer	*21986*

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams (n < 4) en initiale de phrase et de message

Mesure "InitialMsg" : propension des N-grams à apparaître en début de messages (p.r. en initiale de phrase) et mesure du % de messages commençant pas ce N-gram

WikiDisc	InitialMsg	%o Msg
Discussions	99.9	18
Avis	99.7	17
supprimer	97.1	6
conserver	95.7	5
Neutre	98.9	4
Votes	99.8	3
Signalé_par	98.8	3
des_articles_admissibles	99.2	3
Si_vous_êtes	15.6	2
Il_me_semble	32.8	2
Bilan	97.9	2
Il_y_a	23.4	2
Merci	28.2	2
Ce_n'_est	22.8	2
pourBA	98.7	2
Bon	53.6	2
Je_viens_de	60.8	2
Je_ne_vois	32.3	1
Je_suis_d'accord	59.7	1

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams (n < 4) en initiale de phrase et de message

Mesure "InitialMsg" : propension des N-grams à apparaître en début de messages (p.r. en initiale de phrase) et mesure du % de messages commençant pas ce N-gram

WikiDisc	InitialMsg	‰ Msg
Discussions	**99.9**	18
Avis	**99.7**	17
supprimer	**97.1**	6
conserver	**95.7**	5
Neutre	**98.9**	4
Votes	**99.8**	3
Signalé_par	**98.8**	3
des_articles_admissibles	**99.2**	3
Si_vous_êtes	15.6	2
Il_me_semble	32.8	2
Bilan	**97.9**	2
Il_y_a	23.4	2
Merci	28.2	2
Ce_n'_est	22.8	2
pourBA	**98.7**	2
Bon	53.6	2
Je_viens_de	60.8	2
Je_ne_vois	32.3	1
Je_suis_d'accord	59.7	1

- si InitialMsg > 85

$\Rightarrow$ **"code interne Wiki"**

- si InitialMsg < 85

$\Rightarrow$ Argumentation et Recherche d'un consensus

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams ($n < 4$) en initiale de phrase et de message

forum Santé	Inital Msg	‰ Msg
Coucou_les_filles	99.0%	25
coucou_les_filles	99.3%	14
Coucou	89.0%	8
Salut_les_filles	99.2%	7
Bonjour	87.9%	7
Bonjour_les_filles	99.0%	7
coucou	92.0%	5
salut_les_filles	99.0%	4
bonjour_les_filles	99.3%	3
Coucou_les	99.0%	3
Merci	47.4%	2
Merci_les_filles	80.0%	2
bonjour	90.9%	2
Salut	89.2%	2
Bonsoir_les_filles	99.3%	2
coucou_les	99.5%	2
Oui	51.4%	2
Oui_c'_est	57.2%	2
merci_les_filles	85.3%	2
Comment_allez_vous	44.0%	2
merci	57.6%	2
oui_c'_est	66.7%	1

Formules d'ouverture : WikiDiscussions vs. Forum Santé

Méthode : Extraction des n-grams ($n < 4$) en initiale de phrase et de message

forum Santé	Inital Msg	‰ Msg
Coucou_les_filles	**99.0%**	25
coucou_les_filles	**99.3%**	14
Coucou	**89.0%**	8
Salut_les_filles	**99.2%**	7
Bonjour	**87.9%**	7
Bonjour_les_filles	**99.0%**	7
coucou	**92.0%**	5
salut_les_filles	**99.0%**	4
bonjour_les_filles	**99.3%**	3
Coucou_les	**99.0%**	3
Merci	47.4%	2
Merci_les_filles	80.0%	2
bonjour	**90.9%**	2
Salut	**89.2%**	2
Bonsoir_les_filles	**99.3%**	2
coucou_les	**99.5%**	2
Oui	51.4%	2
Oui_c'_est	57.2%	2
merci_les_filles	**85.3%**	2
Comment_allez_vous	44.0%	2
merci	57.6%	2
oui_c'_est	66.7%	1

- Spécificité des locuteurs
- Conversation

Premières conclusions

Première caractérisation du genre "WikiDiscussion"

- Des discussions "bien écrites"
- A priori peu chargées d'*affects*
- A la recherche d'un consensus
- Avec encore pas mal de "code interne Wiki" (à nettoyer ?)

Premières conclusions

Première caractérisation du genre "WikiDiscussion"

- Des discussions "bien écrites"
- A priori peu chargées d'*affects*
- A la recherche d'un consensus
- Avec encore pas mal de "code interne Wiki" (à nettoyer ?)

Approche data-driven : découvrir les caractéristiques des WikiDiscussions

- Poids de N-grams lexicaux, morphologiques, syntaxiques
- dans une tâche classification automatique supervisée

Premières conclusions

Première caractérisation du genre "WikiDiscussion"

- Des discussions "bien écrites"
- A priori peu chargées d'*affects*
- A la recherche d'un consensus
- Avec encore pas mal de "code interne Wiki" (à nettoyer ?)

Approche data-driven : découvrir les caractéristiques des WikiDiscussions

- Poids de N-grams lexicaux, morphologiques, **syntaxiques**
- dans une **tâche classification automatique supervisée**

Approche data-driven : découvrir les traits typiques du genre discussion

Caractériser le web sans se limiter à un domaine particulier

- Inconvénients des approches lexicales classiques, e.g. Scott & Tribble 2006 :
  - Reflètent et dépendent de la thématique des textes
  - Peu utiles si la thématique change

$\Rightarrow$ S'abstraire du domaine et de la thématique des textes

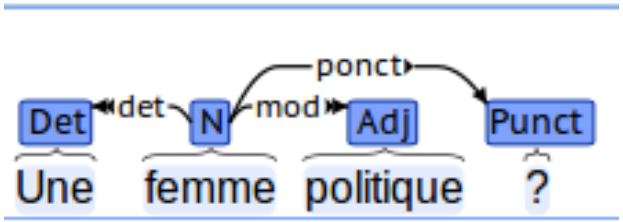
- Pour identifier des caractéristiques génériques aux genres du web

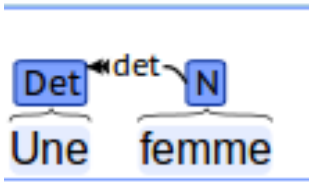
Méthode délexicalisée basée sur des n-grams syntaxiques

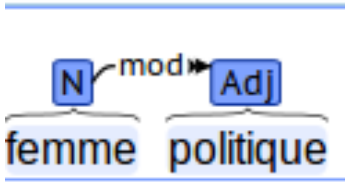
- Observer et illustrer les N-grams typiques dans les WikiDiscussions et les corpus de comparaison
- Évaluation via une tâche de classification

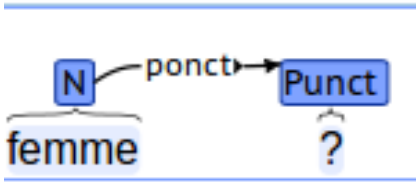
N-grams syntaxiques

- Petits sous-arbres d'une analyse syntaxique en dépendance
- Suivent les relations de dépendance ⇒ pas nécessairement linéaires
- Méthode appliquée à l'origine à l'anglais (Goldberg & Orwant 2013)
- Adaptation au finnois pour le Finnish Internet Parsebank + publication du pipeline (Kanerva & al. 2014)
- Adaptation au français (Laippala & Ho-Dac 2015)



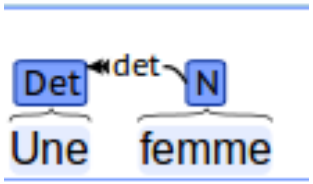


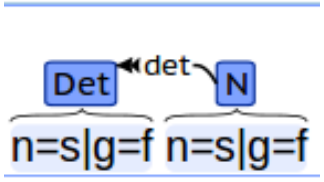


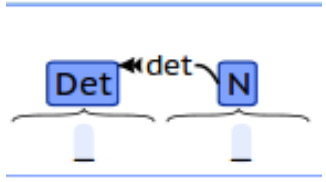


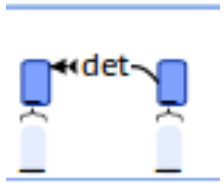
N-grams syntaxiques : paramétrage des niveaux de granularité et de lexicalisation

Conservation des lexèmes, traits morphologiques, classes morphologiques, relations de dépendances







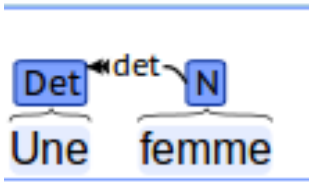


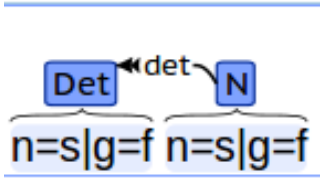
Analyse détaillée
Qté d'informations ++

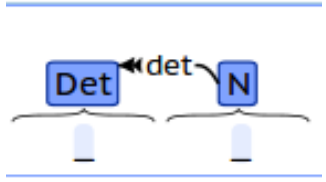
[-] Analyse abstraite
Qté d'information - -

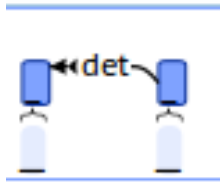
N-grams syntaxiques : paramétrage des niveaux de granularité et de lexicalisation

Conservation des lexèmes, traits morphologiques, classes morphologiques, relations de dépendances









Analyse détaillée
Qté d'informations ++

[-] Analyse abstraite
Qté d'information - -

Pour nous : Méthode d'analyse au-delà du niveau lexical pour faire émerger (classer selon) des structures spécifiques (et non des termes ou des collocations spécifiques)

Méthode de détermination des traits typiques au genre "WikiDiscussion"

Classification automatique WikiDiscussion vs. corpus de comparaison

- corpus d'apprentissage : 50% du corpus
- corpus de test : 50% restants
- segment textuel à classer : 3 phrases
- classifieur : Vowpal Wabbit (Argawal et al. 2011)
- tâche : classer les textes dans la catégorie WikiDiscussion ou pas

Données pour interprétation

- F-mesure pour évaluer l'efficacité de N-grams syntaxiques
- Traits typiques associés par le modèle statistique et générés en sortie

Échantillon du corpus utilisé

Pour réduire le coût du traitement (*work in progress*)

WikiDiscussion ↔ Forum Santé

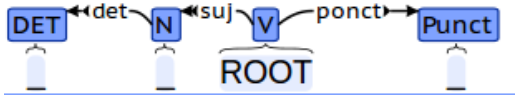
- WikiDiscussion (12 903 816 tokens, 636 553 phrases)
- Forum Santé (12 182 582 tokens, 1 170 791 phrases)

WikiDiscussion ↔ Articles Wikipedia

- WikiDiscussion (4 634 209 tokens, 232 807 phrases)
- Articles Wikipedia (4 349 085 tokens, 212 826 phrases)

N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

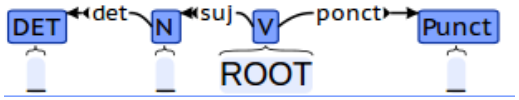
WikiDiscussion F-mesure 91%

Forte présence du tri-arc* 

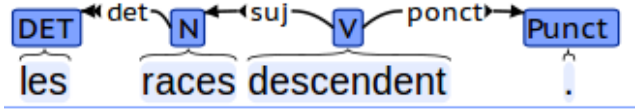
* (avec des traits morphologiques différents)

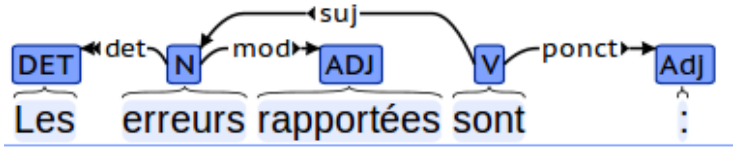
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

WikiDiscussion F-mesure 91%

Forte présence du tri-arc* 

Exemples de contextes qui impliquent cet N-gram :



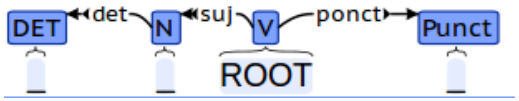


* (avec des traits morphologiques différents)

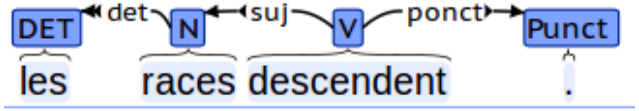
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

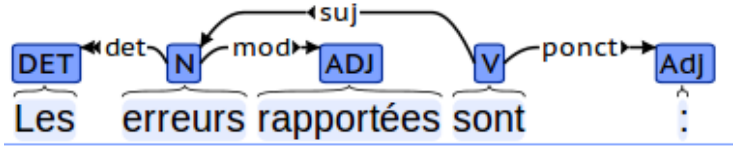
WikiDiscussion F-mesure 91%

Forte présence du tri-arc*

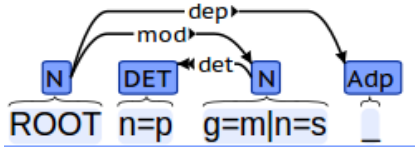


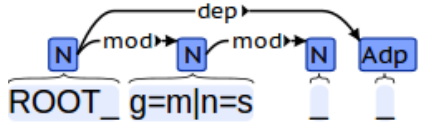
Exemples de contextes qui impliquent cet N-gram :





Signatures (code wiki) :



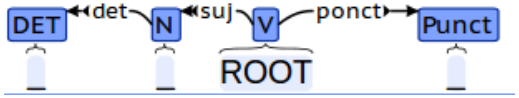


* (avec des traits morphologiques différents)

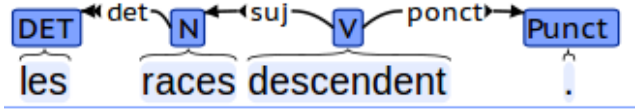
N-grams syntaxiques : tri-arcs avec DEP + POS + MORPHO

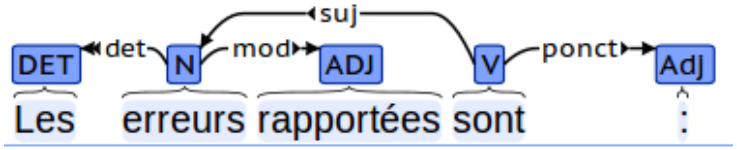
WikiDiscussion F-mesure 91%

Forte présence du tri-arc*

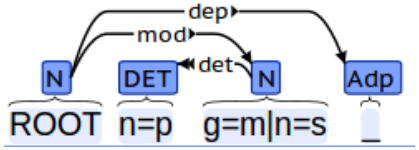


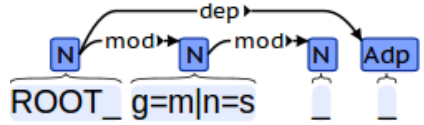
Exemples de contextes qui impliquent cet N-gram :

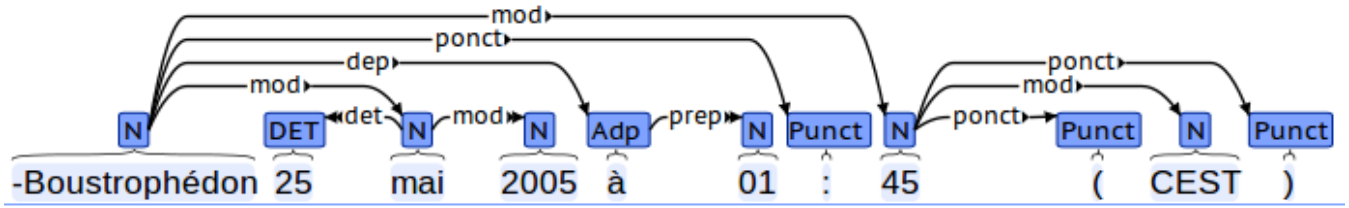




Signatures (code wiki) :



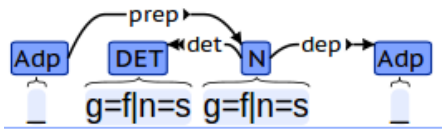


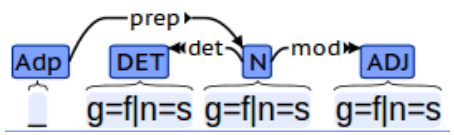


* (avec des traits morphologiques différents)

WikiDiscussion (vs. Forum Santé), F-mesure 91%

Syntagmes prépositionnels :

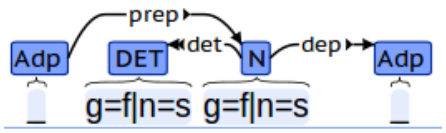


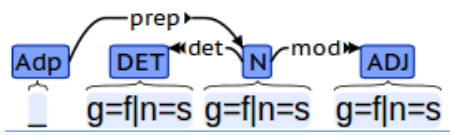


- *... de la défense sans...*
- *... dans leur version initiale...*

WikiDiscussion (vs. Forum Santé), F-mesure 91%

Syntagmes prépositionnels :

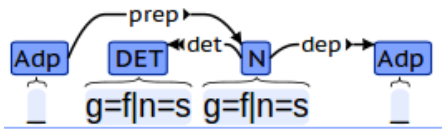


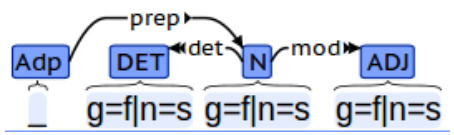


- *... de la défense sans...*
- *... dans leur version initiale...*

WikiDiscussion (vs. Forum Santé), F-mesure 91%

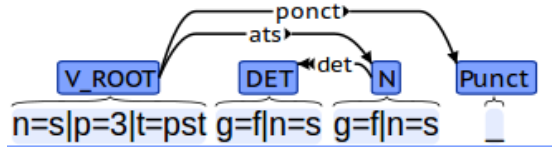
Syntagmes prépositionnels :





- *... de la défense sans...*
- *... dans leur version initiale...*

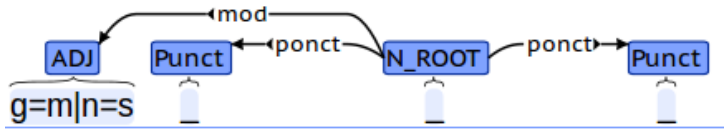
Construction *est*+NOM :



- *... est pas une bonne idée...*
- *... c'est une légende...*

Forum Santé F-mesure 94%

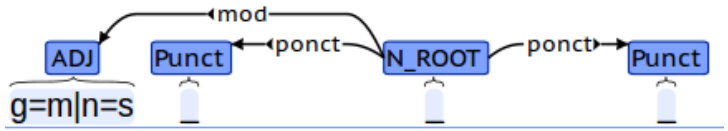
Salutations :



- *Cher (e) Angélique...*

Forum Santé F-mesure 94%

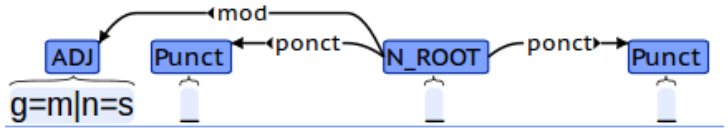
Salutations :



- *Cher (e) Angélique...*

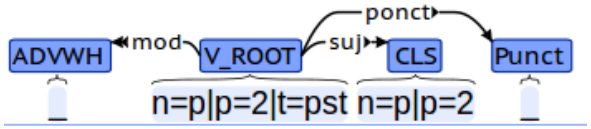
Forum Santé F-mesure 94%

Salutations :



- *Cher (e) Angélique...*

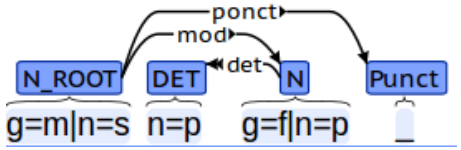
Questions :



- *Pourquoi pleurez-vous ?*

Forum Santé F-mesure 94%

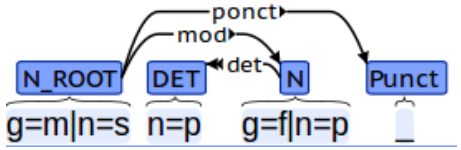
Souhaits :



- *Merci les poulettes ! ! !*
- *Courage les filles !*

Forum Santé F-mesure 94%

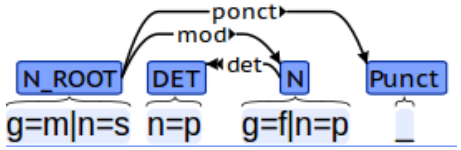
Souhaits :



- *Merci les poulettes ! ! !*
- *Courage les filles !*

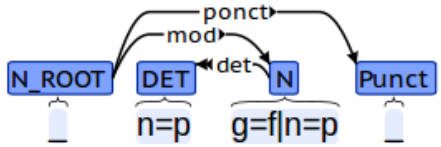
Forum Santé F-mesure 94%

Souhaits :



- *Merci les poulettes ! ! !*
- *Courage les filles !*

Émoticones :



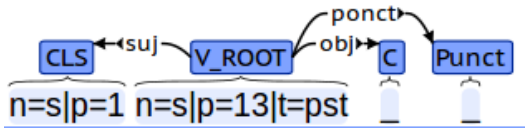
- *EMOTICONE, tes chats ne changent pas...*

Conclusion sur WikiDiscussion vs. Forum Santé

- Les deux corpus montrent des caractéristiques communes : noms d'utilisateurs, phrases nominales (typiques des discussions ?)
- ... mais on voit également des différences très claires !
- Les N-grams des WikiDiscussion sont associés à des phrases complètes (même complexes)
- .... tandis que les N-grams du Forum de Santé sont associés à des souhaits, des salutations familières et des emoticones

Wikipedia discussions (vs. articles), F-mesure 86%

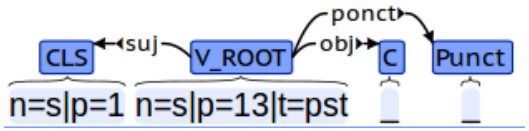
Je+verbe+que :



- *J'ajoute que...*
- *Je pense plutôt que...*

Wikipedia discussions (vs. articles), F-mesure 86%

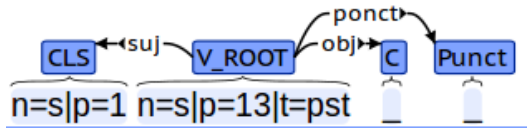
Je+verbe+que :



- *J'ajoute que...*
- *Je pense plutôt que...*

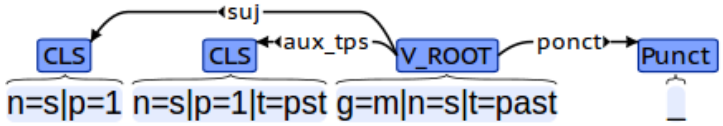
Wikipedia discussions (vs. articles), F-mesure 86%

Je+verbe+que :



- *J'ajoute que...*
- *Je pense plutôt que...*

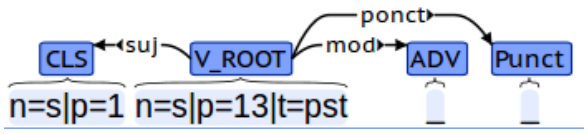
Je+aux+verbe :



- *J'en ai profité...*
- *J'ai retiré...*

Wikipedia discussions (vs. articles), F-mesure 86%

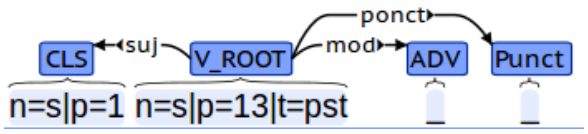
Je+verbe+ADV :



- *Je me retrouve donc...*
- *Je ne pense pas...*
- *Je suis également...*

Wikipedia discussions (vs. articles), F-mesure 86%

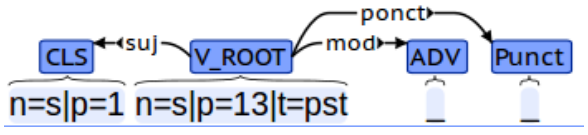
Je+verbe+ADV :



- *Je me retrouve donc...*
- *Je ne pense pas...*
- *Je suis également...*

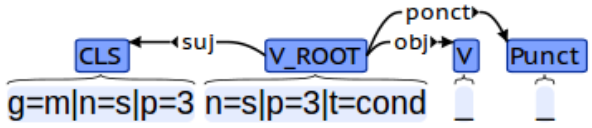
Wikipedia discussions (vs. articles), F-mesure 86%

Je+verbe+ADV :



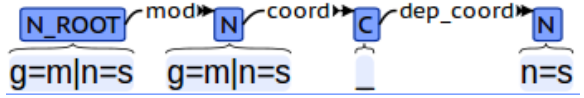
- *Je me retrouve donc...*
- *Je ne pense pas...*
- *Je suis également...*

CLS + cond. → Opinions ! :



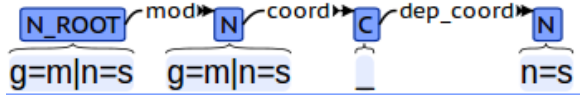
- *Il vaudrait mieux s'en tenir...*
- *Il faudrait expliciter...*

articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

Coordinations dans une phrase nominale : 

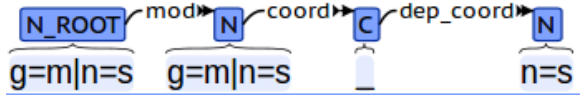
- *Francois d'Aguilon (1567-1617), mathématicien et architecte...*

articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

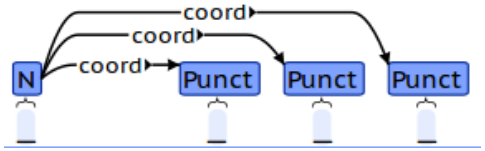
Coordinations dans une phrase nominale : 

- *Francois d'Aguilon (1567-1617), mathématicien et architecte...*

articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

Coordinations dans une phrase nominale : 

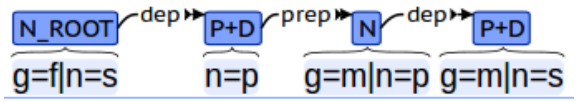
- *Francois d'Aguilon (1567-1617), mathématicien et architecte...*

Autres coordinations : 

- *Villars-Brandis, Taloire, Eoulx...*

articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

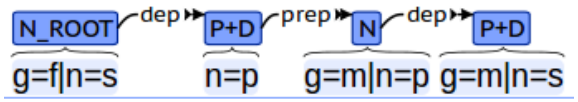
Phrases nominales annonçant une liste :



- *Liste des joueurs du Club...*
- *Liste des députés du Bas-Rhin :*

articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

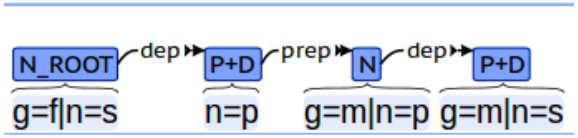
Phrases nominales annonçant une liste :



- *Liste des joueurs du Club...*
- *Liste des députés du Bas-Rhin :*

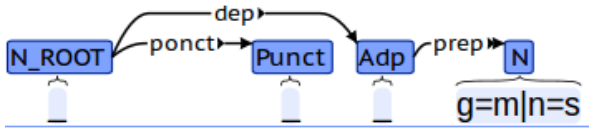
articles Wikipedia (vs. wikiDiscussion), F-mesure 85%

Phrases nominales annonçant une liste :



- *Liste des joueurs du Club...*
- *Liste des députés du Bas-Rhin :*

Phrases nominales explicatives :



- *TDS : En France, Traitement...*

Résultats

WikiDiscussion ↔ Articles Wikipedia

- WikiDiscussion : combinaisons différentes de *je* + *verbe* et *il* + *cond.*
  - → Modalité / subjectivité / argumentation (à creuser)
- Articles Wikipedia : coordinations, phrases nominales
  - → Annonces des débuts des listes, des listes
  - → Constructions visant à expliquer + informer

WikiDiscussion ↔ Forum Santé

- WikiDiscussion : combinaisons différentes de *det+suj+verbe*, syntagmes prépositionnels et *est* + *verbe*
  - → Phrases complètes
- Forum Santé : souhaits, salutations, emoticones

Caractéristiques des WikiDiscussions basées sur les N-grams syntaxiques

- Pas de traits de langage familier !
- Pas de traits de signes d'interaction
- N-grams syntaxiques reflètent plutôt une syntaxe complète
- Traces de subjectivité / argumentation

Et l'utilité des n-grams syntaxiques ?

Petite comparaison aux trigrammes lexicaux

WikiDiscussions, F-mesure 88%

(-cest-) (-cet-) "'-pierre-"' de-l'-article dans-l'-article **il-me-semble ,-j'-ai** l'-article-. **je-n'-ai je-pense-que que-j'-ai ,-je-ne** dans-cet-article je-viens-de de-la-page c'-est-pas que-c'-est **je-ne-sais** il-y-a cet-article-. **je-ne-vois je-pense-qu'** lien-externe-mort **je-suis-d'accord ,-non- ?** je-l'-ai ,-c'-est

Articles Wikipedia F-mesure 89%

avions-actuels-# (-né-en (-ambassade-) ,-gallimard-, danseur-et-chorégraphe (-bretagne-) selon-la-liste artiste-professionnel-# ,-né-à festival-de-cannes)-,-peintre (-danseur-et écrivain-amateur-# liste-des-sénateurs communauté-de-communes)—royaume-uni liste-des-députés (-éteint-))-,-écrivain)-,-français prix-campbell-# ,-coll-. liste-des-préfets artiste-amateur-# #-ee-siècle

N-grams syntaxiques $\leftrightarrow$ lexicaux

WikiDiscussion

- Présence des traces de subjectivité avec les deux méthodes
- Présence des phrases complètes uniquement avec les N-grams syntaxiques

Articles de Wikipedia

- Présence des coordinations, listes, explications uniquement avec les N-grams syntaxiques

Bilan de l'approche par N-grams syntaxique

	Wikidiscussions	**Forum Santé**
N-grams syntaxiques	91%	94%
Bag-of-words	*100%*	*99%*
	Wikidiscussions	**Articles Wikipédia**
N-grams syntaxiques	86%	85%
Trigrammes lexicaux	88%	89%
Bag-of-lemmata	*92%*	*92%*

Conclusions provisoires des analyses data-dirven

- → Les N-grams syntaxiques moyen de classification relativement efficace
- → Certaines caractéristiques reflétées par des N-grams syntaxiques aussi présentes dans les N-grams lexicaux
- → Certaines caractéristiques structurales complètement absentes des N-grams lexicaux
- → Les N-grams syntaxiques sont utiles dans la description des textes
- → ... et permettent une analyse au-delà du niveau lexical !

Plan

Conclusions générales

Sur la caractérisation du genre "WikiDiscussions"

- Des discussions "bien écrites" (peu de mots inconnus et des phrases complètes)
- Un niveau de langue plutôt standard (non familier)
- A priori peu d'*affects* mais de la modalisation et de l'implication du locuteur (je+V)

Sur la complémentarité des approches

- Les résultats semblent se compléter

To Dos

- Simplifier la méthode et éviter le biais du classifieur (e.g. log-likelihood ratio)
- Optimiser les paramètres pour pouvoir décrire la totalité du corpus
  - Sélection des traits (en relation avec les analyses hypothesis-driven)
  - Sélection des discussions "longues"
- Tirer partie des méta-données
  - Observer les variations entre thématiques (e.g. portail physique vs. portail people)
  - Observer les spécificités des discussions "appel au calme"
  - Décrire plus en détail le rôle des différents locuteurs
- Contraster avec un forum de discussion différent (forum Ubuntu ?)

Références

- Agarwal, A., Chappelle, O., Dudik, M. and Langford, J. (2011). A Reliable Effective Terascale Linear Learning System. JMLR, 15, 1111-1133.
- Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., & Rynck, F. (2008) Constitution de ressources pédagogiques numériques : le lexique des affects. Dans M. Loiseau, M. Abouzaïd, L. Buson, C. Cavalla, A. Djaroun, C. Dugua, et al. (éD.), Autour Des Langues Et Du Langage : Perspective Pluridisciplinaire (p. 407–414). Grenoble : Presses Universitaires de Grenoble.
- Gayral, F., Jacques, M.-P., Poibeau, T. and Zimina, M. (2007) Typologie textuelle : état de l'art et applications. Rapport du projet RNTL TEXTCOOP, LIPN, Paris.
- Goldberg, Y., and Orwant, J. (2013). A Dataset of Syntactic-N grams over Time from a Very Large Corpus of English Books. Second Joint Conference on Lexical and Computational Semantics (*SEM), 1. Association for Computational Linguistics.
- Kanerva, J., Luotolahti, J., Laippala, V., and Ginter, F. (2014). Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. Proceedings of the Sixth International Conference Baltic HLT.
- Kilgarriff, A., Reddy, S., Pomikálek, J. and PVS A. (2010). A Corpus Factory for Many Languages. In LREC workshop on Web Services and Processing Pipelines, Malta, May 2010.
- Laippala, Veronika ; Kanerva, Jenna ; Ginter, Filip. 2015. Syntactic Ngrams as Keystructures Reflecting Typical Syntactic Patterns of Corpora in Finnish. Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics. 198, 233-241.
- Laippala, Veronika ; Kanerva, Jenna ; Pyysalo, Sampo ; Missilä, Anna ; Salakoski, Tapio ; Ginter, Filip. 2015. Syntactic N-grams in the Classification of the Finnish Internet Parsebank : Detecting Translations and Informality. Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015), May 11–13, 2015 in Vilnius, Lithuania.
- Scott, M., and Tribble, C. (2006). Textual Patterns : Key Words and Corpus Analysis in Language Education . Philadelphia, PA, USA : John Benjamins Publishing Company.
- Trosborg, A. (1997) Text typology : Register, genre and text type. In Trosborg (ed.) Text typology and translation, John Benjamins, Amsterdam, 3–23.
- Urieli, A. (2013) Analyse syntaxique robuste du français : conciliar methods syntaxiques et connaissances linguistiques dans l'outil Talismane, Thèse de doctorat, Toulouse 2